CHAPTER 14
ITEM ANALYSES

As noted in Brown (1983), "a test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each question. Both the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices in Education* include standards for identifying quality questions. Questions should assess only knowledge or skills that are under assessment and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that MEA questions meet these standards. Previous sections in this report have delineated the qualitative checks on question quality. The current chapter focuses on more quantitative evaluations. The statistical evaluations are presented in three sections: (1) difficulty indices, (2) item-test correlations, and (3) subgroup differences in item performance. The results presented in this chapter are based on the statewide administration of MEA in March of 1998. A total of 16,898 grade 4 students, 17,599 grade 8 students, and 14,547 grade 11 students participated in the assessment.

DIFFICULTY INDICES

All multiple-choice, short-answer, and open-response questions were evaluated in terms of difficulty and relationship to overall score according to standard classical test theory practice. Difficulty was measured by averaging the proportion of points received across all students who received the question. Multiple-choice and short-answer questions were scored dichotomously (correct v. incorrect), so for these questions, the difficulty index is simply the proportion of students who correctly answered the question. Open-response questions allowed for scores between zero and four. By computing the difficulty index as the average proportion of points received, the indices for multiple-choice, short-answer, and open-response questions are placed on a similar

scale; the index ranges from zero to one regardless of the question type. Although this index is traditionally described as a measure of difficulty (as it is described here), it is properly interpreted as an easiness index because larger values indicate easier questions. An index of zero indicates that no student received credit for the question, and an index of one indicates that every student received full credit for the question.

Questions that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. Similarly, questions that are correctly answered by very few students may indicate knowledge or skills that have not yet been mastered by most students, but such questions provide little information about differences in student ability. In general, to provide best measurement, difficulty indices should range from near-chance performance (0.25 for four-option, multiple-choice questions or essentially zero for short-answer and open-response questions) to 0.90. Indices outside this range indicate questions that were either too difficult or too easy for the target population.

ITEM-TEST CORRELATIONS

Although difficulty is an important question characteristic, the relationship between performance on a question and performance on the whole test or a relevant test section may be more critical. A question that assesses relevant knowledge or skills should relate to other questions that are purported to be measuring the same knowledge or skills.

Within classical test theory, these relationships are assessed using correlation coefficients that are typically described as either item-test correlations or, more commonly, discrimination indices. The discrimination index used to analyze MEA multiple-choice items and dichotomous short-answer items was the point-biserial correlation between item score and a criterion total score on the test. Item test correlations for dichotomous items are limited by the difficulty of the item. As such, the index has a theoretical range from –1 to 1, with the magnitude and sign of the index indicating the relationship's strength and direction, respectively. For open-response items, item discrimination indices were based on the Pearson product-moment correlation.

In general, discrimination indices are interpreted as indicating the degree to which high- and low-ability students perform differently on a question or, equivalently, the degree to which performance on a question helps to differentiate between high- and low-ability students. From this perspective, indices near 1 indicate that high-ability students are more likely to answer the question correctly, indices near –1 indicate that low-ability students are more likely to answer the question correctly, and indices near 0 indicate that performance on the question is equally likely to be answered correctly by high- and low-ability students.

Discrimination indices can be thought of as measures of how closely a question assesses the same knowledge and skills assessed by other questions contributing to the criterion total score; that is, the discrimination index can be interpreted as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the MEA, appropriate criterion scores were selected based on item type and function (common or matrix). The selected criterion scores are provided in Table 14-1. For example, the criterion score for common open-response and short-answer items was the total score on all common multiple-choice, open-response, and short-answer items.

Table 14-1
Criterion Score Used in Computing the Discrimination Index
For Each Item Type and Function

Item Type	Item Function	Scores Included in the Total			
		Multiple Choice Common	Multiple Choice Matrix	Open Response Common	Open Response Matrix
Multiple Choice	Common	✓			
	Matrix	✓	✓		
Open Response	Common	✓		✓	
	Matrix	✓	✓	✓	✓

SUMMARY OF ITEM ANALYSIS RESULTS

Frequency distributions and summary statistics of the difficulty and discrimination indices for each question are provided in Tables 14-2–14-4. In general, the question difficulty and discrimination indices are in acceptable and expected ranges. Very few questions were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most questions were assessing consistent constructs, and students who performed well on individual questions tended to perform well overall. There were a small number of questions with near-zero discrimination indices, but none were reliably negative. In a standards-based test, good items may have low or negative item total correlations if schools have not yet aligned their curricula to state curriculum standards.

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the questions or students were common across groups. However, one can say that with respect to multiple-choice items the fourth- and eighth-grade students tended to have more difficulty answering the mathematics questions on the fourth- and eighth-grade tests as compared to the eleventh-grade students answering the math questions on the eleventh-grade tests. In science, the opposite of this statement is true in that eleventh-grade students had more difficulty answering multiple-choice science questions on the eleventh-grade tests than did the other two grades in their respective science tests. The fourth-grade students may have had a slightly easier time with the reading questions on the fourth-grade tests as compared to the eighth-grade students taking the eighth-grade reading questions. Similarly, eighth-grade students may have had a slightly easier time with the reading questions on the eighth-grade tests as compared to the eleventh-grade students taking the eleventh-grade reading questions.

Comparisons within grade levels are reasonable with one caveat: in comparing common and matrix questions, one assumes that the sampling scheme for matrix questions ensures that the students receiving a particular matrix question are representative of the entire population that received the common questions. With that caveat in mind, there appear to be immaterial differences in the mean difficulty of common and matrix multiple-choice questions

regardless of grade level and subject. The exceptions to this observation are the Social Studies multiple-choice questions for grade 4 (with mean difference of 0.13) and grade 11 (with mean difference of 0.11) and Science & Technology grade 11 (with mean difference of 0.17).

Comparing the difficulty indices of multiple-choice and short-answer or open-response questions is inappropriate because multiple-choice questions can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice questions tend to be higher (indicating easier questions) than the difficulty indices for other question types. Similarly, the partial credit allowed by open-response questions is advantageous in the computation of question-test correlations, so the discrimination indices for these questions tend to be larger than the discrimination indices of other question types.

Table 14-2
Average Difficulty and Discrimination of Different Question Types
For Each Subject: Grade 4

Subject	Statistics	Multiple-Choice			Short Answer	Constructed Response
		Common	Matrix	All		
Reading	*n*	23	144	167	39	20
	Average Difficulty	.72	.65	.66	.64	.45
	Average Discrimination	.29	.30	.30	.44	.56
Mathematics	*n*	15	128	143	18	20
	Average Difficulty	.66	.61	.62	.64	.36
	Average Discrimination	.30	.28	.28	.48	.55
Science	*n*	15	128	143	19	23
	Average Difficulty	.58	.58	.58	.42	.34
	Average Discrimination	.20	.20	.20	.39	.48
Social Studies	*n*	15	128	143	15	21
	Average Difficulty	.70	.57	.59	.47	.42
	Average Discrimination	.20	.20	.20	.37	.54
Health	*n*		128		13	27
	Average Difficulty		.75		.88	.49
	Average Discrimination		.18		.39	.58
Visual and Performing Arts	*n*		128		16	31
	Average Difficulty		.56		.45	.42
	Average Discrimination		.15		.43	.62

Table 14-3 Average Difficulty and Discrimination of Different Question Types For Each Subject: Grade 8							
Subject	Statistics	Multiple-Choice			Short Answer	Constructed Response	Extended Response
		Common	Matrix	All			
Reading	*n*	23	144	167	39	20	
	Average Difficulty	.64	.68	.68	.67	.54	
	Average Discrimination	.23	.28	.27	.45	.63	
Mathematics	*n*	15	128	143	19	19	17
	Average Difficulty	.53	.54	.54	.46	.33	.34
	Average Discrimination	.35	.30	.31	.55	.62	.74
Science	*n*	15	128	143	19	43	4
	Average Difficulty	.54	.55	.55	.39	.36	.41
	Average Discrimination	.18	.20	.20	.35	.56	.67
Social Studies	*n*	15	128	143	19	53	
	Average Difficulty	.59	.56	.57	.46	.18	
	Average Discrimination	.21	.24	.23	.38	.61	
Health	*n*		128		16	17	14
	Average Difficulty		.71		.68	.50	.46
	Average Discrimination		.15		.34	.67	.82
Visual and Performing Arts	*n*		128		14	46	
	Average Difficulty		.56		.30	.41	
	Average Discrimination		.15		.42	.72	

Table 14-4 Average Difficulty and Discrimination of Different Question Types For Each Subject: Grade 11							
Subject	Statistics	Multiple-Choice			Short Answer	Constructed Response	Extended Response
		Common	Matrix	All			
Reading	*n*	23	144	167	39	20	
	Average Difficulty	.74	.74	.74	.64	.54	
	Average Discrimination	.29	.28	.28	.45	.62	
Mathematics	*n*	15	128	143	20	18	17
	Average Difficulty	.49	.45	.45	.28	.31	.28
	Average Discrimination	.28	.28	.28	.52	.68	.77
Science	*n*	22	128	150	19	47	2
	Average Difficulty	.31	.48	.46	.61	.34	.37
	Average Discrimination	.13	.20	.19	.41	.61	.71
Social Studies	*n*	15	128	143	16	21	21
	Average Difficulty	.52	.39	.41	.28	.33	.27
	Average Discrimination	.16	.15	.15	.44	.63	.74
Health	*n*		128		12	16	14
	Average Difficulty		.79		.62	.49	.48
	Average Discrimination		.17		.29	.65	.85
Visual and Performing Arts	*n*		128		10	46	
	Average Difficulty		.57		.58	.42	
	Average Discrimination		.14		.47	.77	

SUBGROUP DIFFERENCES IN QUESTION PERFORMANCE

The *Code of Fair Testing Practices in Education* explicitly states that subgroup differences in performance should be examined when sample sizes permit, and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* includes similar guidelines. As part of the effort to identify such problems, MEA questions were evaluated in terms of differential item functioning (DIF) statistics.

DIF procedures are designed to identify questions for which subgroups of interest perform differently beyond the impact of differences in overall achievement. For the MEA, the standardization DIF procedure (Dorans and

Kulick, 1986) was employed to evaluate subgroup differences between male and female. This procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score, then an overall average is calculated weighting the total score distribution so it is the same for the two groups.

The index ranges from –1 to 1 for multiple-choice and short-answer questions and is adjusted to the same scale for open-response questions. Negative numbers indicate that the question was more difficult for female students. Positive numbers indicate that the question was easier for female students.

Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible for dichotomously scored questions (such as MEA multiple-choice questions). Most MEA questions fall within this range. Dorans and Holland further stated that dichotomously scored questions with values between –0.10 and –0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that questions with values outside the [–0.10, 0.10] range (i.e., "high" DIF) are more unusual and should be examined very carefully. These standards can be applied to open-response questions by accounting for the larger range of possible index values and scaling appropriately. That is, values of the DIF index can range from –4.0 to 4.0, so the corresponding ranges are between –0.2 and 0.2 for negligible difference, between –0.4 and –0.2 and between 0.2 and 0.4 for "low" DIF and outside [-0.4, 0.4] for "high" DIF.

DIF indices indicate differential performance between two groups. That differential performance may or may not be indicative of bias in the test. Course-taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup differences in performance are related to construct-relevant factors, the questions should be considered for inclusion on a test.

Each question was categorized according to the guidelines adapted from Dorans and Holland (1993). Tables 14-5 to 14-7 provide the number of questions in each of the three DIF categories for male v. female for each grade level tested. There are some MEA questions categorized as "low" or "high" DIF. These indices must not be

interpreted as indisputable evidence of bias. Both the *Code of Fair Testing Practices in Education* and the *Standards for Educational and Psychological Testing* assert that test questions must be free from construct-irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct-relevant factors, the questions may be included on a test. What is important is to determine if the cause of this differential performance is construct relevant.

Table 14-5 Differential Item Functioning (DIF) Categorization of Common Items by Item Type: Grade 4													
Subject	Item Type	"High" DIF				"Low" DIF				Negligible DIF			
		Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%
Reading	Multiple Choice	0	0	0	0.00	0	1	1	2.94	11	11	22	64.71
	Short Answer	0	0	0	0.00	1	0	1	2.94	2	4	6	17.65
	Constructed Response	0	0	0	0.00	0	0	0	0.00	4	0	4	11.76
	All Items	0	0	0	0.00	1	1	2	5.88	17	15	32	94.12
Mathematics	Multiple Choice	0	1	1	4.35	1	2	3	13.04	5	6	11	47.83
	Short Answer	0	0	0	0.00	0	0	0	0.00	2	1	3	13.04
	Constructed Response	0	0	0	0.00	0	0	0	0.00	4	1	5	21.74
	All Items	0	1	1	4.35	1	2	3	13.04	11	8	19	82.61
Science	Multiple Choice	0	0	0	0.00	0	2	2	8.70	4	9	13	56.52
	Short Answer	0	0	0	0.00	1	0	1	4.35	2	0	2	8.70
	Constructed Response	0	0	0	0.00	0	0	0	0.00	3	2	5	21.74
	All Items	0	0	0	0.00	1	2	3	13.04	9	11	20	86.96
Social Studies	Multiple Choice	0	0	0	0.00	0	5	5	22.73	4	6	10	45.45
	Short Answer	0	0	0	0.00	0	0	0	0.00	0	2	2	9.09
	Constructed Response	0	0	0	0.00	0	0	0	0.00	5	0	5	22.73
	All Items	0	0	0	0.00	0	5	5	22.73	9	8	17	77.27

Table 14-6
Differential Item Functioning (DIF) Categorization of Common Items by Item Type: Grade 8

Subject	Item Type	"High" DIF				"Low" DIF				Negligible DIF			
		Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%
Reading	Multiple Choice	0	0	0	0.00	1	4	5	15.15	7	10	17	51.52
	Short Answer	0	0	0	0.00	0	0	0	0.00	6	1	7	21.21
	Constructed Response	0	0	0	0.00	0	0	0	0.00	4	0	4	12.12
	All Items	0	0	0	0.00	1	4	5	15.15	17	11	28	84.85
Mathematics	Multiple Choice	0	0	0	0.00	0	5	5	22.73	4	6	10	45.45
	Short Answer	0	0	0	0.00	0	0	0	0.00	1	2	3	13.64
	Constructed Response	0	0	0	0.00	1	0	0	4.55	1	1	2	9.09
	Extended Response	0	0	0	0.00	0	0	0	0.00	0	1	1	4.55
	All Items	0	0	0	0.00	1	5	6	27.27	6	10	16	72.73
Science	Multiple Choice	0	2	2	8.70	4	2	6	26.09	0	7	7	30.43
	Short Answer	0	0	0	0.00	1	0	1	4.35	1	2	3	13.04
	Constructed Response	1	0	0	4.35	0	0	0	0.00	2	1	3	13.04
	All Items	1	2	3	13.04	5	2	7	30.43	3	10	13	56.52
Social Studies	Multiple Choice	0	2	2	8.70	0	2	2	8.70	4	7	11	47.83
	Short Answer	0	0	0	0.00	0	0	0	0.00	0	3	3	13.04
	Constructed Response	0	0	0	0.00	2	0	2	8.70	3	0	3	13.04
	All Items	0	2	2	8.70	2	2	4	17.39	7	10	17	73.91

Table 14-7 Differential Item Functioning (DIF) Categorization of Common Items by Item Type: Grade 11													
Subject	Item Type	"High" DIF				"Low" DIF				Negligible DIF			
		Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%	Favor Female	Favor Male	N	%
Reading	Multiple Choice	0	0	0	0.00	0	2	2	5.88	15	6	21	61.76
	Short Answer	0	0	0	0.00	0	0	0	0.00	6	1	7	20.59
	Constructed Response	0	0	0	0.00	1	0	1	2.94	3	0	3	8.82
	All Items	0	0	0	0.00	1	2	3	8.82	24	7	31	91.18
Mathematics	Multiple Choice	0	0	0	0.00	0	1	1	4.55	2	12	14	63.64
	Short Answer	0	0	0	0.00	1	0	1	4.55	1	1	2	9.09
	Constructed Response	0	0	0	0.00	0	1	1	4.55	1	1	2	9.09
	Extended Response	0	0	0	0.00	0	0	0	0.00	0	1	1	4.55
	All Items	0	0	0	0.00	1	2	3	13.64	4	15	19	86.36
Science	Multiple Choice	0	0	0	0.00	0	0	0	0.00	3	12	15	68.18
	Short Answer	0	1	1	4.55	0	1	1	4.55	0	1	1	4.55
	Constructed Response	0	0	0	0.00	0	0	0	0.00	2	1	3	13.64
	Extended Response	0	0	0	0.00	1	0	1	4.55	0	0	0	0.00
	All Items	0	1	1	4.55	1	1	2	9.09	5	14	19	86.36
Social Studies	Multiple Choice	0	0	0	0.00	0	1	1	4.55	6	8	14	63.64
	Short Answer	0	0	0	0.00	0	2	2	9.09	0	0	0	0.00
	Constructed Response	0	0	0	0.00	1	0	1	4.55	4	0	4	18.18
	All Items	0	0	0	0.00	1	3	4	18.18	10	8	18	81.82